ASSESS CONSCIOUSNESS FROM LONG-DURATION EEG MONITORING VIA EVIDENCE ACCUMULATION TRANSFORMER

Anonymous authors
Paper under double-blind review

ABSTRACT

Assessing consciousness in patients with severe disorders of consciousness plays a critical role in diagnosis, prognosis, rehabilitation planning, and clinical decision-making. Reliable EEG-based assessment requires long-duration monitoring to capture consciousness-related patterns that may be sparse, intermittent, and evolving. Identifying nuanced signatures of consciousness from long-duration EEG remains challenging because evidence must be localized and accumulated across the recording, while supervision is available only through subject-level diagnoses. We propose EviFormer, the Evidence Accumulation Transformer, to assess consciousness from long-duration EEG. EviAttention applies attention along time and across electrodes, then models temporal patterns of different durations through a mixture of experts. EviMemory uses gated causal convolutions to accumulate recurring evidence across EEG chunks. EviLearning supervises subject-level evidence seeking by learning representations over evidence sets. We collect a sleep EEG dataset involving 140 subjects spanning different levels of consciousness, including 60 subjects diagnosed with disorders of consciousness. Across the collected cohorts and public datasets, EviFormer achieves state-of-the-art or competitive performance, while attribution analyses highlight delta-band and temporal evidence together with long-range anterior–posterior and bilateral electrode patterns.

1 INTRODUCTION

Disorders of consciousness are severe acquired brain injury syndromes that impair wakefulness and awareness. We refer to them as DoC. Chronic patients may be diagnosed with unresponsive wakefulness syndrome, which we abbreviate as UWS; minimally conscious state, or MCS; or emergence from MCS, or EMCS (Carriere et al., 2022). This distinction informs prognosis, rehabilitation, monitoring, and treatment decisions. Diagnosis nevertheless depends largely on eliciting and interpreting behavioral responses, which requires specialized expertise and repeated observations (Giacino et al., 2018; Wannez et al., 2017). Motor impairment, low arousal, and fluctuating responsiveness can conceal preserved cognition, so a single examination may miss intermittent signs of consciousness. Repeated, standardized assessment is therefore important for reliable longitudinal monitoring (Giacino et al., 2018; Schnakers et al., 2009; Wannez et al., 2017; Cruse et al., 2011).

We use the abbreviation EEG for electroencephalography. EEG provides a direct and objective window on residual brain function even when a patient cannot produce a reliable motor response (Cruse et al., 2011; Sitt et al., 2014). It has high temporal resolution, measures neural activity non-invasively in real time, and is suitable for bedside monitoring, thereby enabling more objective and timely clinical assessment. Studies have consequently identified several EEG markers of consciousness. Sleep organization and sleep spindles are preserved more consistently in MCS than in UWS (Landsness et al., 2011). Low-frequency power, signal complexity, and inter-electrode information exchange discriminate consciousness level (Sitt et al., 2014), while connectivity hubs correlate with behavioral diagnosis, brain metabolism, and prognosis (Chennu et al., 2017). These findings provide a physiological basis for data-driven DoC modeling.

As EEG datasets and computational methods have advanced, deep learning has become a major approach to EEG decoding, with architectures including recurrent neural networks, convolutional

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

neural networks, graph neural networks, and Transformers (Lawhern et al., 2018; Supratak et al., 2017; Song et al., 2023). More recently, LaBraM, EEGPT, CBraMod, and CSBrain have learned reusable representations from heterogeneous EEG corpora and achieved competitive or state-of-the-art transfer across downstream tasks (Jiang et al., 2024; Wang et al., 2024a; 2025; Zhou et al., 2025). Representation learning has also supported increasingly automated DoC assessment from spontaneous EEG, TMS-evoked EEG, P300 responses, and other stimulus-evoked activity (Lee et al., 2022; Qu et al., 2026; Pan et al., 2023; Cai et al., 2025).

Moving from short-stimulus EEG decoding to long-term monitoring poses an unexpected challenge. Simply applying a seconds-scale EEG decoder to more chunks independently and averaging its predictions discards temporal dependencies and leaves unclear how intermittent evidence forms a patient-level judgment. This issue is especially important in DoC because neural signs of residual consciousness may be sparse, intermittent, and evolving (Wannez et al., 2017; Landsness et al., 2011; Della Bella et al., 2025). This raises two problems. First, evidence accumulation over long recordings must connect recurring and evolving evidence so that prolonged observation can reinforce the patient-level judgment. Second, patient diagnoses without chunk labels create a weakly supervised chunk-modeling problem.

Long-duration DoC detection data are scarce because few public datasets and limited numbers of available subjects make data-driven model exploration more difficult. We first collect a clinical sleep EEG dataset of 140 participants spanning the spectrum of consciousness and cognition, including a core DoC cohort of 60 patients. We then draw on language-model advances in long-sequence processing and memory (Beltagy et al., 2020; Gu & Dao, 2023; Yang et al., 2025; Wang et al., 2024d; Behrouz et al., 2025), and propose the Evidence Accumulation Transformer, or EviFormer, to address evidence accumulation in long-record EEG. EviEmbedding represents short-term evidence within each EEG chunk through temporal and spectral views. EviAttention applies attention along time and across electrodes, then models temporal patterns of different durations through a mixture of experts. EviMemory uses gated causal convolutions to accumulate recurring evidence across EEG chunks. EviLearning supervises subject-level evidence seeking by learning representations over evidence sets without assigning diagnostic labels to individual chunks. We evaluate EviFormer on our self-collected cohorts and public EEG datasets. Under the reported within-cohort protocols, its mean BAcc and AUROC are the highest among the listed methods on the two clinical endpoints. A separate context-budget diagnostic examines when longer observation helps. Evidence analyses associate delta and alpha activity and temporal, parietal, and occipital electrodes with model decisions; these sensor-level associations are not causal biomarkers.

2 Related Work

Traditional EEG-based DoC assessment combines handcrafted spectral, complexity, connectivity, and stimulus-response features with patient-level classifiers (Sitt et al., 2014; Liang et al., 2020; Engemann et al., 2018; Toppi et al., 2024; Wang et al., 2024b; 2022). More recently, deep models have achieved state-of-the-art performance by learning nonlinear representations directly from EEG. Cai et al. pretrained a self-supervised contrastive domain-generalization network on music-evoked EEG from healthy participants and transferred it to decode stimulus responses in patients with DoC (Cai et al., 2025). Qu et al. interpolated channel-wise approximate entropy from resting and preferred-music EEG into topographic maps and trained a CNN to distinguish UWS from MCS (Qu et al., 2026). While these models make progress on stimulus responses in patients with DoC, DoC-Informer recently combined parallel temporal convolutions, electrode-position-aware spatial attention, temporal attention, and channel-masking training to classify UWS and MCS under resting-state EEG, moving toward bedside monitoring (Yang et al., 2026).

3 Method

We propose EviFormer to assess consciousness from long-duration EEG. It first represents local evidence within each EEG chunk, then models this evidence over different physical time ranges, and finally accumulates chunk representations as the recording unfolds. Training proceeds in two consecutive stages. The model first learns a local encoder from unlabeled EEG and then trains the full sequence model directly using subject-level labels.

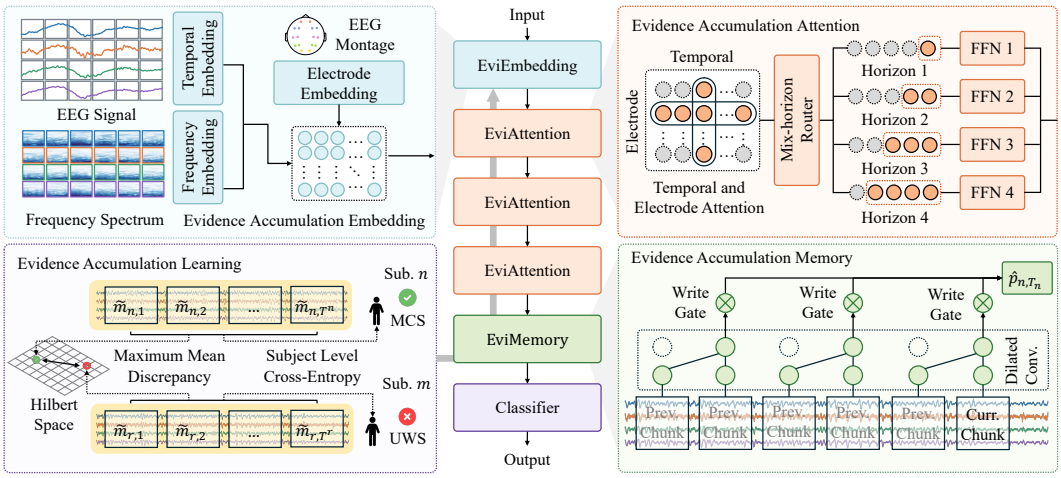


Figure 1: The proposed EviFormer framework. EviEmbedding represents short-term EEG evidence through temporal and spectral views. EviAttention identifies patterns over different time windows within each chunk by combining temporal and electrode attention with feed-forward experts. EviMemory uses gated causal convolutions to accumulate recurring evidence across chunks. EviLearning optimizes the aggregated subject prediction and subject-level evidence distributions without assigning labels to individual chunks.

3.1 EviFormer Architecture

Let $\mathcal{D} = \{(\mathbf{X}_n, y_n)\}_{n=1}^{N}$ denote a labeled dataset. The sequence $\mathbf{X}_n = (\mathbf{x}_{n,1}, \ldots, \mathbf{x}_{n,T_n})$ contains T_n ordered EEG chunks for subject n, and $y_n \in \{0, 1\}$ is the subject label. Each chunk is denoted by $\mathbf{x}_{n,t} \in \mathbb{R}^{C \times L}$, where C is the number of electrodes and L is the number of samples in one chunk.

EviEmbedding. EviEmbedding splits each electrode trace in a chunk into patches of P samples; adjacent patches overlap by $P/2$ samples. Let $\mathbf{u}_{n,t,c,p} \in \mathbb{R}^P$ denote patch p of electrode c, with $p = 0, \ldots, Q-1$, where Q is the number of patches per electrode. We encode each patch through waveform and spectral branches. The resulting view embeddings are

$$\begin{aligned} \mathbf{a}^{\text{temp}}_{n,t,c,p} &= \mathbf{W}_{\text{temp}} \,\text{Pool}\left(\text{Conv}_{\text{temp}}(\mathbf{u}_{n,t,c,p})\right), \\ \mathbf{a}^{\text{freq}}_{n,t,c,p} &= \mathbf{W}_{\text{freq}} \log\left(1 + |\text{rFFT}(\mathbf{u}_{n,t,c,p})|\right). \end{aligned} \tag{1}$$

In the waveform branch, $\text{Conv}_{\text{temp}}$ consists of two one-dimensional temporal convolutions with 32 output channels. The first uses kernel width 25 and stride 5, and the second uses kernel width 5 and stride 1. Each is followed by group normalization and GELU; Pool averages the remaining time positions. In the spectral branch, the forward-normalized one-sided real FFT returns $P/2 + 1$ frequency coefficients. We take their magnitudes and apply $\log(1 + x)$ elementwise. Linear maps $\mathbf{W}_{\text{temp}}$ and $\mathbf{W}_{\text{freq}}$ project the resulting waveform and spectral vectors to dimension d.

We concatenate the two projected views and map them back to d dimensions. To identify where the patch was recorded, $\mathbf{e}_c \in \mathbb{R}^d$ is a learned vector indexed by electrode identity. A fixed sinusoidal vector $\mathbf{p}_p \in \mathbb{R}^d$ is computed from the patch start time within the chunk. The resulting patch token is

$$\mathbf{H}^{(0)}_{n,t,c,p} = \mathbf{W}_{\text{fuse}} \left[\mathbf{a}^{\text{temp}}_{n,t,c,p}; \mathbf{a}^{\text{freq}}_{n,t,c,p}\right] + \mathbf{e}_c + \mathbf{p}_p. \tag{2}$$

EviAttention. EviAttention seeks evidence distributed across different temporal scales and electrodes within a long-duration EEG chunk. At each stage s, it takes a token grid $\mathbf{H}^{(s)}$ and applies two parallel axial-attention operations. Temporal attention treats the time positions of each electrode as items; electrode attention treats the C electrodes at each time position as items. Mathematically, the

parallel electrode and temporal attention operations for one head are

$$\mathrm{Attn}^{(s)}_{\mathrm{temp}}(\mathbf{H})_{c,:,:} = \mathrm{softmax}\left(\frac{\mathbf{Q}^{(s)}_{c,:,:}(\mathbf{K}^{(s)}_{c,:,:})^\top}{\sqrt{d_h}}\right)\mathbf{V}^{(s)}_{c,:,:},$$
$$\mathrm{Attn}^{(s)}_{\mathrm{elec}}(\mathbf{H})_{:,p,:} = \mathrm{softmax}\left(\frac{\mathbf{Q}^{(s)}_{:,p,:}(\mathbf{K}^{(s)}_{:,p,:})^\top}{\sqrt{d_h}}\right)\mathbf{V}^{(s)}_{:,p,:}, \tag{3}$$

where $\mathbf{Q}^{(s)}$, $\mathbf{K}^{(s)}$, and $\mathbf{V}^{(s)}$ are learned projections and d_h is the head dimension. Outputs from all heads are concatenated and projected in the usual way. Learned residual gates combine both axes with the incoming representation according to

$$\widetilde{\mathbf{H}}^{(s)} = \mathbf{H}^{(s)} + \tanh(\alpha^{(s)}_{\mathrm{temp}})\,\mathrm{Attn}^{(s)}_{\mathrm{temp}}(\mathrm{LN}(\mathbf{H}^{(s)})) + \tanh(\alpha^{(s)}_{\mathrm{elec}})\,\mathrm{Attn}^{(s)}_{\mathrm{elec}}(\mathrm{LN}(\mathbf{H}^{(s)})). \tag{4}$$

The fused axial-attention output $\widetilde{\mathbf{H}}^{(s)}$ then enters a mixture of experts. Expert e corresponds to a physical window of r_e seconds and first averages neighboring tokens along time within its window, with stride one and zero padding at the boundaries, producing one window vector for each original token position. Its two-layer feed-forward network maps $d \to 4d \to d$ with GELU between the linear layers. Expert outputs and routing probabilities are

$$\mathbf{E}^{(s)}_e(\widetilde{\mathbf{H}}^{(s)})_{c,p} = \mathrm{FFN}^{(s)}_e\left(\mathrm{AvgPool}_{r_e}(\mathrm{LN}(\widetilde{\mathbf{H}}^{(s)}))_{c,p}\right),$$
$$\boldsymbol{\pi}^{(s)}_{c,p} = \mathrm{softmax}\left(\mathbf{W}^{(s)}_r\,\mathrm{LN}(\widetilde{\mathbf{H}}^{(s)}_{c,p}) + \mathbf{b}^{(s)}_r\right), \qquad \mathcal{K}^{(s)}_{c,p} = \mathrm{Top2}(\boldsymbol{\pi}^{(s)}_{c,p}). \tag{5}$$

The affine router $(\mathbf{W}^{(s)}_r, \mathbf{b}^{(s)}_r)$ scores all four experts separately at each electrode and time position and selects the two highest-probability experts, $\mathcal{K}^{(s)}_{c,p}$. Their probabilities are renormalized and used to combine their outputs. A learned residual gate then adds the resulting multi-duration update $\mathbf{M}^{(s)}$ to the attention output as follows.

$$\mathbf{M}^{(s)}_{c,p} = \sum_{e\in\mathcal{K}^{(s)}_{c,p}} \frac{\pi^{(s)}_{c,p,e}}{\sum_{j\in\mathcal{K}^{(s)}_{c,p}}\pi^{(s)}_{c,p,j}}\mathbf{E}^{(s)}_e(\widetilde{\mathbf{H}}^{(s)})_{c,p},$$
$$\widehat{\mathbf{H}}^{(s)} = \widetilde{\mathbf{H}}^{(s)} + \tanh(\alpha^{(s)}_{\mathrm{moe}})\mathbf{M}^{(s)}. \tag{6}$$

We stack J EviAttention blocks to capture DoC-related evidence in the EEG recording. Let $\mathbf{F}_{n,t} \in \mathbb{R}^{C\times Q\times d}$ denote the resulting fused token grid. Layer normalization, averaging over electrodes and time, and a final linear map yield the following chunk representation.

$$\mathbf{z}_{n,t} = \mathbf{W}_o\,\mathrm{Mean}_{c,p}\left(\mathrm{LN}(\mathbf{F}_{n,t,c,p})\right) \in \mathbb{R}^D. \tag{7}$$

EviMemory. EviMemory integrates evidence across chunks and accumulates it over the observed history. Let $\mathbf{v}_{n,t}$ be its adapted input and set $\mathbf{h}^{(0)}_{n,t} = \mathbf{v}_{n,t}$. We stack two three-tap depthwise causal convolution blocks with dilations $d_1 = 1$ and $d_2 = 2$. A dilation of d_ℓ spaces neighboring taps d_ℓ chunks apart, so block ℓ reads positions t, $t - d_\ell$, and $t - 2d_\ell$ without using future information. In this way, each historical observation is enriched with contextual information from preceding chunks. After layer normalization and causal convolution, a pointwise convolution produces content $\mathbf{b}$ and modulation $\mathbf{c}$. $\sigma(\mathbf{c})$ modulates $\mathbf{b}$ featurewise, followed by dropout and a linear projection to produce the candidate update.

$$[\mathbf{b}^{(\ell)}_{n,t}; \mathbf{c}^{(\ell)}_{n,t}] = \mathrm{PWConv}^{(\ell)}\left(\mathrm{CConv}_{d_\ell}\left(\mathrm{LN}(\mathbf{h}^{(\ell-1)}_{n,\le t})\right)\right),$$
$$\mathbf{u}^{(\ell)}_{n,t} = \mathbf{W}^{(\ell)}_o\left(\mathbf{b}^{(\ell)}_{n,t} \odot \sigma(\mathbf{c}^{(\ell)}_{n,t})\right). \tag{8}$$

Here, PWConv, CConv, and LN denote pointwise convolution, dilated causal depthwise convolution, and layer normalization. Because not all historical evidence contributes equally to the task, a two-layer write gate compares the normalized current and preceding representations and controls how much of the candidate update is written into the current memory representation.

$$\mathbf{g}^{(\ell)}_{n,t} = \sigma\left(\mathrm{Gate}^{(\ell)}\left[\mathrm{LN}(\mathbf{h}^{(\ell-1)}_{n,t}); \mathrm{LN}(\mathbf{h}^{(\ell-1)}_{n,t-1})\right]\right),$$
$$\mathbf{h}^{(\ell)}_{n,t} = \mathbf{h}^{(\ell-1)}_{n,t} + \tanh(\alpha^{(\ell)}_{\mathrm{mem}})\mathbf{g}^{(\ell)}_{n,t} \odot \mathbf{u}^{(\ell)}_{n,t}. \tag{9}$$

Each block passes its residual output to the next block. For the first chunk, the preceding state is zero. Let $\overline{\mathbf{h}}_{n,t}$ be the output of the second block and $\boldsymbol{\delta}_{n,t} = \overline{\mathbf{h}}_{n,t} - \mathbf{v}_{n,t}$ its change relative to the current adapted chunk vector. We bound that change relative to $\|\mathbf{v}_{n,t}\|_2$ using $c_{\text{mem}} = 0.5$ as follows.

$$\mathbf{m}_{n,t} = \mathbf{v}_{n,t} + \min\left(1, \frac{c_{\text{mem}}\|\mathbf{v}_{n,t}\|_2}{\|\boldsymbol{\delta}_{n,t}\|_2 + \epsilon}\right)\boldsymbol{\delta}_{n,t}. \tag{10}$$

The bound prevents history from overwhelming the current chunk. We accumulate evidence over the full observed history by averaging the memory-enhanced vectors through time t and applying a linear classifier.

$$\mathbf{s}_{n,t} = \frac{1}{t}\sum_{j=1}^{t}\mathbf{m}_{n,j}, \qquad \widehat{p}_{n,t} = \sigma(\mathbf{w}^\top \mathbf{s}_{n,t} + b), \qquad \widehat{p}_n = \widehat{p}_{n,T_n}. \tag{11}$$

3.2 EviLearning

Pretraining. Pretraining updates EviEmbedding and EviAttention without the memory or subject classifier. From one unlabeled recording, we pair a current chunk $\mathbf{x}_{i,t}$ with a future chunk $\mathbf{x}_{i,t+\Delta}$ sampled one to eight chunks later. The shared encoder maps them to $\mathbf{z}_i$ and $\mathbf{z}_i^+$. The first objective predicts features extracted from the future chunk by a fixed feature extractor $\overline{\varphi}$. A linear head ρ maps $\mathbf{z}_i$ to those features, with Smooth-L1 loss

$$\mathcal{L}_{\text{future}} = \frac{1}{B}\sum_{i=1}^{B}\ell_{\text{SL1}}\left(\rho(\mathbf{z}_i) - \overline{\varphi}(\mathbf{x}_{i,t+\Delta})\right). \tag{12}$$

The second objective is temporal contrastive learning. The sampled future chunk from the same recording is the positive for each current chunk; the other future chunks in the mini-batch serve as negatives. A two-layer projection head ψ maps both encoder outputs to unit vectors $\mathbf{q}_i = \psi(\mathbf{z}_i)/\|\psi(\mathbf{z}_i)\|_2$ and $\mathbf{k}_i = \psi(\mathbf{z}_i^+)/\|\psi(\mathbf{z}_i^+)\|_2$. With $s_{ij} = \mathbf{q}_i^\top \mathbf{k}_j/\tau$, the symmetric loss is

$$\mathcal{L}_{\text{con}} = -\frac{1}{2B}\sum_{i=1}^{B}\left[\log\frac{\exp(s_{ii})}{\sum_{j=1}^{B}\exp(s_{ij})} + \log\frac{\exp(s_{ii})}{\sum_{j=1}^{B}\exp(s_{ji})}\right]. \tag{13}$$

This loss increases similarity between each current chunk and its paired future chunk relative to the other batch futures. We minimize $\mathcal{L}_{\text{future}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}}$ with $\lambda_{\text{con}} = 0.05$.

Fine-tuning. EviLearning samples one consecutive set of positions in each subject's raw-chunk order. The model processes all selected chunks, averages their memory outputs, and produces one probability $\widehat{p}_n$ for the set. Only this subject-level probability receives the diagnosis label. With class weight w_k, the classification loss is

$$\mathcal{L}_{\text{cls}} = -\frac{1}{N_{\text{tr}}}\sum_{n=1}^{N_{\text{tr}}} w_{y_n}\left[y_n \log\widehat{p}_n + (1-y_n)\log(1-\widehat{p}_n)\right]. \tag{14}$$

Thus, each set contributes one subject target, and the model combines informative and uninformative intervals without assigning that label to individual chunks.

We additionally supervise how evidence is distributed across a subject's memory outputs, rather than only their mean. Each subject's K memory vectors form an empirical distribution. For subjects n and r, the following radial-basis maximum mean discrepancy compares within-subject and between-subject kernel similarities.

$$D_{nr} = \frac{1}{K^2}\sum_{i,j=1}^{K}\kappa(\widetilde{\mathbf{m}}_{n,i}, \widetilde{\mathbf{m}}_{n,j}) + \frac{1}{K^2}\sum_{i,j=1}^{K}\kappa(\widetilde{\mathbf{m}}_{r,i}, \widetilde{\mathbf{m}}_{r,j}) - \frac{2}{K^2}\sum_{i,j=1}^{K}\kappa(\widetilde{\mathbf{m}}_{n,i}, \widetilde{\mathbf{m}}_{r,j}), \qquad \kappa(\mathbf{a},\mathbf{b}) = \exp\left(-\gamma\|\mathbf{a}-\mathbf{b}\|_2^2\right), \tag{15}$$

Table 1: Subject-level comparison on DoC-60 and MMSE-140. DoC-60 distinguishes UWS from MCS, and MMSE-140 distinguishes the MMSE-zero and MMSE-positive groups. We report the mean $\pm$ standard deviation of balanced accuracy and AUC-ROC on the test sets. An asterisk indicates a significant corrected repeated k-fold cross-validation t-test against every comparator for that metric.

	DoC-60		MMSE-140	
Method	BAcc ↑	AUROC ↑	BAcc ↑	AUROC ↑
TCN (Lea et al., 2017)	$.608 \pm .120$	$.635 \pm .148$	$.575 \pm .082$	$.612 \pm .091$
ModernTCN (Luo & Wang, 2024)	$.573 \pm .112$	$.612 \pm .137$	$.568 \pm .080$	$.607 \pm .090$
MedGNN (Fan et al., 2025)	$.541 \pm .110$	$.566 \pm .128$	$.536 \pm .088$	$.558 \pm .101$
iTransformer (Liu et al., 2024)	$.567 \pm .135$	$.605 \pm .148$	$.574 \pm .082$	$.612 \pm .093$
PatchTST (Nie et al., 2023)	$.559 \pm .111$	$.598 \pm .133$	$.563 \pm .084$	$.601 \pm .094$
TimesNet (Wu et al., 2023)	$.554 \pm .127$	$.592 \pm .151$	$.557 \pm .087$	$.590 \pm .099$
TinySleepNet (Supratak & Guo, 2020)	$.579 \pm .137$	$.687 \pm .164$	$.606 \pm .077$	$.629 \pm .081$
DeepSleepNet (Supratak et al., 2017)	$.592 \pm .116$	$.647 \pm .132$	$.596 \pm .079$	$.638 \pm .088$
XSleepNet (Phan et al., 2022)	$.623 \pm .121$	$.640 \pm .136$	$.588 \pm .083$	$.617 \pm .092$
EEGNet (Lawhern et al., 2018)	$.558 \pm .132$	$.589 \pm .124$	$.552 \pm .086$	$.581 \pm .095$
EEGInception (Zhang et al., 2021)	$.601 \pm .118$	$.654 \pm .152$	$.579 \pm .081$	$.625 \pm .089$
EEGConformer (Song et al., 2023)	$.614 \pm .121$	$.672 \pm .150$	$.601 \pm .078$	$.646 \pm .086$
MNet (Watanabe et al., 2024)	$.552 \pm .109$	$.581 \pm .127$	$.541 \pm .085$	$.566 \pm .098$
ADformer (Wang et al., 2024c)	$.586 \pm .130$	$.624 \pm .141$	$.585 \pm .081$	$.624 \pm .091$
DoC-Informer (Yang et al., 2026)	$.601 \pm .113$	$.664 \pm .172$	$.606 \pm .090$	$.616 \pm .117$
LEAD (Wang et al., 2026)	$.620 \pm .099$	$.665 \pm .158$	$.590 \pm .079$	$.641 \pm .087$
CBraMod (Wang et al., 2025)	$.613 \pm .118$	$.680 \pm .143$	$.582 \pm .084$	$.626 \pm .094$
LaBraM (Jiang et al., 2024)	$.641 \pm .124$	$.712 \pm .149$	$.608 \pm .077$	$.655 \pm .087$
CSBrain (Zhou et al., 2025)	$.648 \pm .130$	$.721 \pm .137$	$.612 \pm .075$	$.661 \pm .085$
EviFormer	$\mathbf{.737 \pm .094}^*$	$\mathbf{.761 \pm .132}^*$	$\mathbf{.650 \pm .066}^*$	$\mathbf{.684 \pm .085}^*$

Here $\widetilde{\mathbf{m}}$ is an ℓ_2-normalized memory output. Let $\mathcal{V}$ contain batch subjects with at least one same-class partner, and let η rescale batch distances. We use the following objective to pull same-class evidence distributions together relative to all other subjects.

$$\mathcal{L}_{\text{pop}} = -\frac{1}{|\mathcal{V}|} \sum_{n \in \mathcal{V}} \log \frac{\sum_{r \neq n,\, y_r = y_n} \exp(-\eta D_{nr}/\tau_{\text{pop}})}{\sum_{r \neq n} \exp(-\eta D_{nr}/\tau_{\text{pop}})}. \tag{16}$$

Finally, the trainable MoE router in DoC fine-tuning receives a balancing loss $\mathcal{L}_{\text{bal}} = E \sum_{e=1}^{E} \bar{\pi}_e \bar{s}_e$, where $E = 4$, $\bar{\pi}_e$ is the batch-mean routing probability for expert e, and $\bar{s}_e$ is its batch-mean top-2 selection frequency. This term enters the DoC objective with weight 0.01 and discourages routing collapse.

4 Experiments

4.1 Dataset Description

We collect a bedside-monitoring EEG dataset spanning severe disorders of consciousness and a broader spectrum of consciousness and cognition. We refer to the broader cognitive-assessment cohort as MMSE-140. It contains 140 subjects spanning severe DoC, different degrees of cognitive impairment, and normal cognition. Each subject was assessed with the Mini-Mental State Examination, or MMSE, which is a standardized cognitive screening instrument (Folstein et al., 1975). We refer to the severe DoC cohort as DoC-60. It is a 60-subject subset in which each subject has a clinical consciousness-state diagnosis. It contains 21 subjects with UWS and 39 subjects in the MCS class. The latter combines 18 subjects with MCS$-$, 11 with MCS$+$, and 10 with EMCS.

We use 16 electrodes arranged according to the international 10–20 system. The electrodes are Fp1, Fp2, F3, F4, F7, F8, T3, T4, T5, T6, C3, C4, P3, P4, O1, and O2. We apply a 50 Hz notch filter and a 0.5 to 45 Hz band-pass filter, resample every signal to 100 Hz, and divide it into non-overlapping 30-second chunks. Thus, $C = 16$ and $L = 3000$. We formulate two subject-level binary classification tasks. MMSE-140 distinguishes the MMSE-zero group ($\text{MMSE} = 0$) from the MMSE-positive group ($\text{MMSE} > 0$), assessing whether measurable cognitive function is present; DoC-60 distinguishes the UWS class from the MCS class, addressing the central clinical distinction in severe DoC assessment.

4.2 Comparison Experiments

We evaluate subject-level classification on DoC-60 and MMSE-140. We use ten repetitions of subject-wise five-fold cross-validation. In each repetition, we partition subjects into five folds and use each

fold in turn as the test set, with the other four folds for training. Table 1 groups the baselines by model family. The general time-series models are TCN, ModernTCN, MedGNN, iTransformer, PatchTST, and TimesNet. The task-specific EEG models include the sleep-staging models TinySleepNet, DeepSleepNet, and XSleepNet. EEGNet, EEGInception, and EEGConformer target motor imagery or general EEG decoding, while MNet, ADformer, and DoC-Informer target clinical EEG tasks. The pretrained EEG foundation models are LEAD, CBraMod, LaBraM, and CSBrain.

EviFormer has the highest mean BAcc and AUROC on both tasks in Table 1. Relative to the runner-up on every metric, CSBrain, its BAcc gains are approximately 9 and 4 percentage points on DoC-60 and MMSE-140, respectively; its AUROC gains are approximately 4 and 2 percentage points. We computed fold-wise paired differences between EviFormer and each comparator using identical subject splits. To account for dependence from overlapping training sets and repeated partitions, we used the corrected repeated k-fold cross-validation t-test; significant differences are marked in Table 1. Among the comparison models, the sleep-staging model XSleepNet performs well on DoC-60, while TinySleepNet performs well on MMSE-140. This suggests that sleep models designed around 30-second windows and capable of processing longer sequences are suitable for long-duration EEG monitoring (Phan et al., 2022; Supratak & Guo, 2020). The resting-state DoC model DoC-Informer also performs well, suggesting some consistency across DoC tasks (Yang et al., 2026). The pretrained EEG foundation models CBraMod, LaBraM, and CSBrain also show competitive performance. Their pretraining on clinical EEG data benefits the downstream tasks (Wang et al., 2025; Jiang et al., 2024; Zhou et al., 2025).

4.3 Ablation Studies

We keep the other experimental settings fixed and separately ablate EviEmbedding, EviAttention, EviMemory, and EviLearning over ten repetitions of subject-wise five-fold cross-validation. When ablating EviEmbedding, we use the convolutional layers of EEG Conformer to embed EEG patches and represent local patterns (Song et al., 2023). When ablating EviAttention, we use standard Transformer self-attention (Vaswani et al., 2017). Without EviMemory, we remove the memory module, process all available historical chunks and the current chunk with the ablated model, and average-pool their representations for final classification. Without EviLearning, we fine-tune using the subject label as the chunk label. Check marks indicate retained modules, and boldface marks the highest mean for each metric.

Table 2: Ablations of operations in EviEmbedding, EviAttention, EviMemory, and EviLearning on DoC-60 and MMSE-140.

Embedding	Attention	Memory	Learning	DoC-60		MMSE-140	
				BAcc ↑	AUROC ↑	BAcc ↑	AUROC ↑
	✓	✓	✓	$.647 \pm .145$	$.711 \pm .139$	$.592 \pm .081$	$.622 \pm .096$
✓		✓	✓	$.655 \pm .139$	$.720 \pm .135$	$.601 \pm .079$	$.631 \pm .093$
✓	✓		✓	$.626 \pm .146$	$.704 \pm .143$	$.584 \pm .083$	$.612 \pm .098$
✓	✓	✓		$.638 \pm .144$	$.696 \pm .141$	$.577 \pm .084$	$.603 \pm .100$
✓	✓	✓	✓	$\mathbf{.737 \pm .094}$	$\mathbf{.761 \pm .132}$	$\mathbf{.650 \pm .066}$	$\mathbf{.684 \pm .085}$

The complete model has the highest BAcc and AUROC on both DoC-60 and MMSE-140. Removing any module lowers all four metrics, indicating that each module contributes to performance. The largest losses follow removal of EviMemory or EviLearning, highlighting the value of temporal EEG evidence accumulation and subject-level supervision in both clinical tasks. Removing EviEmbedding or EviAttention also lowers performance, indicating that local representations and multiple temporal horizons improve the evidence available for subject-level decisions.

4.4 Comparison Experiments on Public Datasets

Publicly available EEG datasets with subject-level DoC labels remain scarce. We therefore evaluate EviFormer on public Alzheimer's disease datasets as related cognitive-impairment tasks. Alzheimer's disease is a progressive neurodegenerative disorder characterized by cognitive decline and commonly lower MMSE scores. This experiment examines whether EviFormer's subject-level evidence aggregation is effective across several public EEG datasets beyond our collected cohorts.

Table 3: Subject-level AD-versus-healthy-control comparison on ADFSU, ADFTD-RS, ADFTD-PS, and ADSZ. An asterisk indicates a significant corrected repeated k-fold cross-validation t-test against every comparator for that metric.

	ADFSU		ADFTD-RS		ADFTD-PS		ADSZ	
Method	BAcc ↑	AUROC ↑	BAcc ↑	AUROC ↑	BAcc ↑	AUROC ↑	BAcc ↑	AUROC ↑
TCN	.665 ± .048	.707 ± .057	.588 ± .042	.615 ± .039	.551 ± .045	.559 ± .060	.734 ± .046	.812 ± .047
ModernTCN	.634 ± .065	.709 ± .031	.555 ± .028	.615 ± .032	.528 ± .031	.548 ± .037	.704 ± .029	.802 ± .037
MedGNN	.594 ± .032	.579 ± .063	.508 ± .064	.517 ± .038	.505 ± .051	.515 ± .042	.655 ± .065	.729 ± .060
iTransformer	.542 ± .026	.596 ± .046	.522 ± .036	.556 ± .052	.505 ± .032	.515 ± .064	.637 ± .031	.726 ± .051
PatchTST	.597 ± .027	.684 ± .060	.522 ± .049	.575 ± .049	.512 ± .062	.543 ± .057	.669 ± .058	.778 ± .056
TimesNet	.586 ± .025	.643 ± .025	.534 ± .054	.580 ± .062	.505 ± .060	.532 ± .056	.681 ± .038	.757 ± .037
TinySleepNet	.683 ± .041	.814 ± .027	.589 ± .062	.701 ± .064	.561 ± .056	.628 ± .033	.771 ± .052	.862 ± .047
DeepSleepNet	.683 ± .062	.789 ± .060	.580 ± .036	.665 ± .054	.539 ± .060	.604 ± .047	.774 ± .030	.854 ± .028
XSleepNet	.679 ± .063	.757 ± .042	.573 ± .037	.670 ± .055	.560 ± .054	.558 ± .063	.776 ± .061	.834 ± .044
EEGNet	.645 ± .025	.690 ± .059	.551 ± .052	.621 ± .053	.511 ± .063	.529 ± .041	.725 ± .053	.780 ± .059
EEGInception	.665 ± .064	.763 ± .026	.577 ± .028	.682 ± .058	.537 ± .034	.606 ± .050	.738 ± .041	.861 ± .064
EEGConformer	.664 ± .065	.735 ± .030	.600 ± .040	.643 ± .048	.527 ± .040	.558 ± .050	.767 ± .058	.820 ± .065
MNet	.631 ± .032	.601 ± .041	.574 ± .057	.550 ± .050	.533 ± .038	.515 ± .042	.699 ± .045	.759 ± .036
ADformer	.645 ± .062	.651 ± .049	.585 ± .061	.571 ± .045	.535 ± .042	.515 ± .040	.722 ± .045	.766 ± .030
DoC-Informer	.675 ± .062	.832 ± .074	.589 ± .053	.653 ± .001	.591 ± .010	.639 ± .066	.665 ± .059	.759 ± .075
LEAD	.646 ± .058	.724 ± .054	.594 ± .033	.665 ± .046	.520 ± .053	.565 ± .044	.732 ± .050	.834 ± .030
CBraMod	.633 ± .048	.742 ± .025	.588 ± .041	.666 ± .059	.524 ± .047	.579 ± .054	.728 ± .033	.837 ± .054
LaBraM	.667 ± .045	.774 ± .048	.596 ± .050	.712 ± .049	.554 ± .029	.623 ± .047	.774 ± .032	.857 ± .048
CSBrain	.683 ± .056	.816 ± .036	.615 ± .031	.705 ± .030	.542 ± .044	.606 ± .052	.766 ± .037	.879 ± .058
EviFormer	**.842 ± .020***	**.916 ± .013***	**.775 ± .069***	**.861 ± .035***	**.701 ± .010***	**.737 ± .022***	**.979 ± .021***	**.997 ± .003***

We evaluate Alzheimer's disease versus healthy-control classification on ADFSU, ADFTD-RS, ADFTD-PS, and ADSZ using ten repetitions of subject-wise five-fold cross-validation. ADFSU contains 80 participants with probable Alzheimer's disease and 12 healthy older adults (Vicchietti et al., 2023). After excluding the frontotemporal dementia group, ADFTD-RS and ADFTD-PS each contain 36 participants with Alzheimer's disease and 29 healthy controls (Miltiadous et al., 2023; Ntetska et al., 2025). After excluding the schizophrenia cohort, ADSZ contains 24 participants with Alzheimer's disease and 24 age-matched controls (Alves et al., 2022). The two ADFTD conditions use the same participant splits. We use the comparison methods and evaluation setting described in Section 4.2. Table 3 lists the same methods in the same order as Table 1. We report the mean ± standard deviation of subject-level balanced accuracy and AUROC on held-out test folds. Boldface indicates the highest mean for each metric.

EviFormer achieves the highest mean BAcc and AUROC on all four public datasets. Variation among other methods likely reflects dataset and recording differences and motivates evaluating long-recording evidence across distinct cohorts. Pretrained EEG foundation models remain competitive, while EviFormer improves mean performance across datasets. These results show strong within-dataset performance across four public Alzheimer's disease EEG datasets and suggest that the proposed architecture is applicable to related cognitive-impairment tasks.

4.5 Sensitivity Analyses

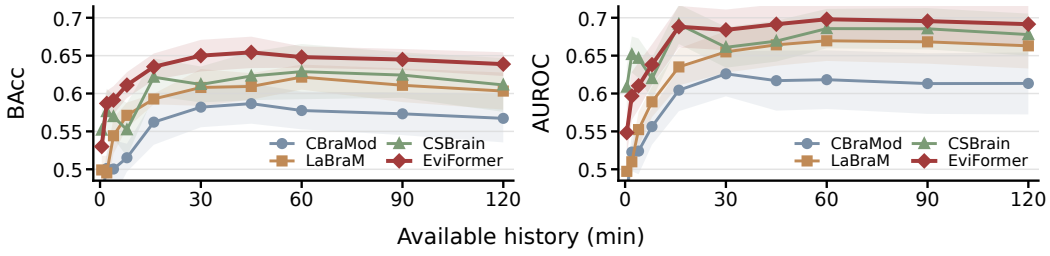
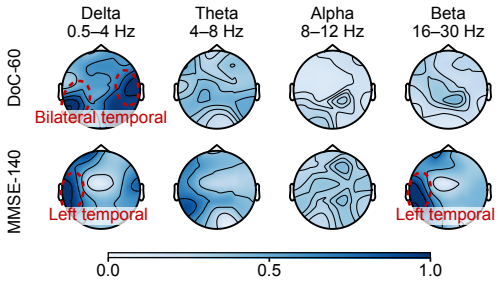


Figure 2: MMSE-140 sensitivity to available EEG history. Shaded bands show the standard error across ten repetitions.

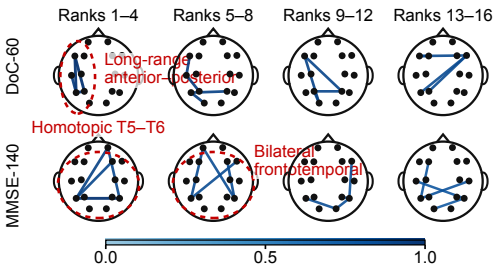
Following the experimental setting in Section 4.2, we analyze sensitivity to the length of EEG history available to the model. We compare EviFormer with CBraMod, LaBraM, and CSBrain at consecutive 30-second observation points. At each point, a model receives the current segment and preceding history up to the specified budget; unavailable history is left-padded and masked. We sweep history budgets from 30 seconds to 120 minutes and evaluate each subject at the terminal point of the

available recording, retaining all 140 subjects. Figure 2 shows rapid gains, followed by saturation and a slight decline. Short histories can miss sparse, intermittent consciousness-related patterns, whereas beyond approximately 30 minutes, additional context is often redundant or noisy. EviFormer accumulates evidence more effectively and outperforms the foundation models after 15 minutes, with 30–60 minutes providing the best performance–duration trade-off.

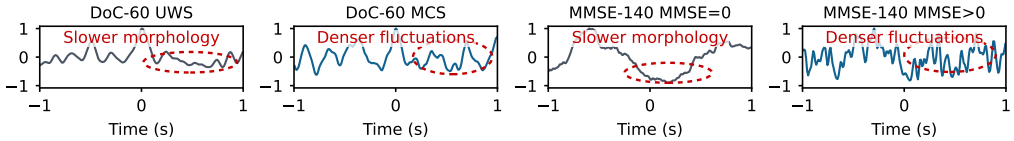
4.6 Qualitative Analyses



(a) Model-attributed evidence is shown from frequency and scalp-region perspectives.



(b) Model-attributed evidence is shown from an electrode-relation perspective.



(c) Model-attributed waveform prototypes are shown as a qualitative case study.

Figure 3: Qualitative model evidence on DoC-60 and MMSE-140.

In this section, we conduct qualitative analyses to visualize the evidence accumulated by EviFormer. For frequency and scalp-region evidence, we visualize the mean change in model output after suppressing one frequency band at one electrode. DoC-60 relies primarily on posterior and temporal delta activity, whereas MMSE-140 emphasizes temporal beta together with frontal and temporal delta. These associations suggest that the model uses temporal slow-wave structure when estimating consciousness level, while temporal beta may capture cognitive or motor-related activity associated with MMSE status. For electrode relations, we calculate spatial-attention matrices, average each electrode-pair value within the cohort, and rank the resulting relations. The results show that the model captures a left-lateralized, long-range anterior–posterior pattern on DoC-60 and bilateral frontotemporal relations on MMSE-140, including the homotopic T5–T6 pair. These patterns may reflect model sensitivity to interhemispheric and anterior–posterior coordination. For the waveform case study, we verify candidate two-second windows by output perturbation and retain those whose removal weakens evidence for the corresponding class. We cluster attributed events by morphology and show the subject-balanced medoid of the dominant cluster for each class. These four prototypes visualize the typical waveforms of the dominant attributed-event cluster in each class. As shown, the dominant UWS and MMSE-zero prototypes contain slower deflections, whereas the MCS and MMSE-positive prototypes exhibit denser fluctuations.

5 Conclusion

We collected a clinical sleep EEG dataset spanning disorders of consciousness and cognition. EviFormer enables long-duration monitoring with multi-scale within-chunk attention, gated cross-chunk memory, and diagnosis-level learning without chunk labels. It achieves state-of-the-art or competitive performance across collected and public datasets. Attribution highlights delta-band, temporal, anterior–posterior, and bilateral electrode evidence, supporting long-recording evidence accumulation. Broader external validation and causal interpretation remain future work.

Reproducibility Statement

Section 3 describes the algorithm, and Section 4 specifies the evaluation setting and reports means and standard deviations. The appendix provides dataset statistics, hyperparameters, and implementation details. The code and anonymized dataset accompany the submission and will be made public after acceptance.

Ethics Statement

The study was approved by the ethics committee of the authors' institution (approval no. 2022(040)). Written informed consent was obtained from all participants or their legally authorized representatives before participation. The study was conducted in accordance with the Declaration of Helsinki.

AI Use Statement

Generative AI assisted with polishing and clarifying prose. The authors reviewed the AI-assisted work and take responsibility for the final content.

References

Caroline L. Alves, Aruane M. Pineda, Kirstin Roster, Christiane Thielemann, and Francisco A. Rodrigues. EEG functional connectivity and deep learning for automatic diagnosis of brain disorders: Alzheimer's disease and schizophrenia. *Journal of Physics: Complexity*, 3(2):025001, 2022.

Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. Titans: Learning to memorize at test time. *Advances in Neural Information Processing Systems*, 38:113506–113543, 2025.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

Honghua Cai, Jiahui Pan, Qiuyi Xiao, Jiarui Jin, Yuanqing Li, and Qiuyou Xie. Decoding musical neural activity in patients with disorders of consciousness through self-supervised contrastive domain generalization. *IEEE Transactions on Affective Computing*, 16(2):726–743, 2025.

Manon Carriere, Roberto Llorens, María Dolores Navarro, Jose Olaya, Joan Ferri, and Enrique Noe. Behavioral signs of recovery from unresponsive wakefulness syndrome to emergence of minimally conscious state after severe brain injury. *Annals of Physical and Rehabilitation Medicine*, 65(2): 101534, 2022.

Srivas Chennu, Jitka Annen, Sarah Wannez, Aurore Thibaut, Camille Chatelle, Helena Cassol, Géraldine Martens, Caroline Schnakers, Olivia Gosseries, David Menon, and Steven Laureys. Brain networks predict metabolism, diagnosis and prognosis at the bedside in disorders of consciousness. *Brain*, 140(8):2120–2132, 2017.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

Damian Cruse, Srivas Chennu, Camille Chatelle, Tristan A. Bekinschtein, Davinia Fernández-Espejo, John D. Pickard, Steven Laureys, and Adrian M. Owen. Bedside detection of awareness in the vegetative state: a cohort study. *The Lancet*, 378(9809):2088–2094, 2011.

Gabriel A. Della Bella, Di Zang, Peng Gui, Diego M. Mateos, Jacobo D. Sitt, Tristan A. Bekinschtein, Dragana Manasova, Benjamine Sarton, Fabrice Ferré, Stein Silva, Pedro W. Lamberti, Xuehai Wu, Ying Mao, Liping Wang, and Pablo Barttfeld. Detection of EEG dynamic complex patterns in disorders of consciousness. *Communications Biology*, 8:1204, 2025.

Denis A. Engemann, Federico Raimondo, Jean-Rémi King, Benjamin Rohaut, Gilles Louppe, Frédéric Faugeras, Jitka Annen, Helena Cassol, Olivia Gosseries, Diego Fernandez-Slezak, Steven Laureys, Lionel Naccache, Stanislas Dehaene, and Jacobo D. Sitt. Robust EEG-based cross-site and cross-protocol classification of states of consciousness. *Brain*, 141(11):3179–3192, 2018.

Wei Fan, Jingru Fei, Dingyu Guo, Kun Yi, Xiaozhuang Song, Haolong Xiang, Hangting Ye, and Min Li. Towards multi-resolution spatiotemporal graph learning for medical time series classification. In *Proceedings of the ACM on Web Conference 2025*, pp. 5054–5064, 2025.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Marshal F. Folstein, Susan E. Folstein, and Paul R. McHugh. "Mini-mental state": A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12(3): 189–198, 1975.

Joseph T. Giacino, Douglas I. Katz, Nicholas D. Schiff, John Whyte, Eric J. Ashman, Stephen Ashwal, Richard Barbano, Flora M. Hammond, Steven Laureys, et al. Practice guideline update recommendations summary: Disorders of consciousness. *Neurology*, 91(10):450–460, 2018.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Weibang Jiang, Liming Zhao, and Bao-liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *International Conference on Learning Representations*, 2024.

Eric Landsness, Marie-Aurélie Bruno, Quentin Noirhomme, Brady Riedner, Olivia Gosseries, Caroline Schnakers, Marcello Massimini, Steven Laureys, Giulio Tononi, and Mélanie Boly. Electrophysiological correlates of behavioural changes in vigilance in vegetative state and minimally conscious state. *Brain*, 134(8):2222–2232, 2011.

Vernon J. Lawhern, Amelia J. Solon, Nicholas R. Waytowich, Stephen M. Gordon, Chou P. Hung, and Brent J. Lance. EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5):056013, 2018.

Colin Lea, Michael D. Flynn, René Vidal, Austin Reiter, and Gregory D. Hager. Temporal convolutional networks for action segmentation and detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 156–165, 2017.

Minji Lee, Leandro R. D. Sanz, Alice Barra, Audrey Wolff, Jaakko O. Nieminen, Melanie Boly, Mario Rosanova, Silvia Casarotto, Olivier Bodart, Jitka Annen, Aurore Thibaut, Rishabh Panda, Vincent Bonhomme, Marcello Massimini, Giulio Tononi, Steven Laureys, Olivia Gosseries, and Seong-Whan Lee. Quantifying arousal and awareness in altered states of consciousness using interpretable deep learning. *Nature Communications*, 13:1064, 2022.

Zhenhu Liang, Shuting Shao, Zhen Lv, Dong Li, Jamie W. Sleigh, Xiaoli Li, Cheng Zhang, and Jianghong He. Constructing a consciousness meter based on the combination of non-linear measurements and genetic algorithm-based support vector machine. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(2):399–408, 2020.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations*, 2024.

Donghao Luo and Xue Wang. ModernTCN: A modern pure convolution structure for general time series analysis. In *International Conference on Learning Representations*, 2024.

Andreas Miltiadous, Katerina D. Tzimourta, Theodora Afrantou, Panagiotis Ioannidis, Nikolaos Grigoriadis, Dimitrios G. Tsalikakis, Pantelis Angelidis, Markos G. Tsipouras, Euripidis Glavas, Nikolaos Giannakeas, and Alexandros T. Tzallas. A dataset of scalp EEG recordings of Alzheimer's disease, frontotemporal dementia and healthy subjects from routine EEG. *Data*, 8(6):95, 2023.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

Aimilia Ntetska, Andreas Miltiadous, Markos G. Tsipouras, Katerina D. Tzimourta, Theodora Afrantou, Panagiotis Ioannidis, Dimitrios G. Tsalikakis, Konstantinos Sakkas, Emmanouil D. Oikonomou, Nikolaos Grigoriadis, Pantelis Angelidis, Nikolaos Giannakeas, and Alexandros T. Tzallas. A complementary dataset of scalp EEG recordings featuring participants with Alzheimer's disease, frontotemporal dementia, and healthy controls, obtained from photostimulation EEG. *Data*, 10(5):64, 2025.

Jiahui Pan, Honghua Cai, Haiyun Huang, Yanbin He, and Yuanqing Li. Multiple scale convolutional few-shot learning networks for online P300-based brain–computer interface and its application to patients with disorder of consciousness. *IEEE Transactions on Instrumentation and Measurement*, 72:1–16, 2023.

Bo Peng, Daniel Goldstein, Quentin Anthony, Alon Albalak, Eric Alcaide, Stella Biderman, Eugene Cheah, Xingjian Du, Teddy Ferdinan, Haowen Hou, et al. Eagle and finch: Rwkv with matrix-valued states and dynamic recurrence. *arXiv preprint arXiv:2404.05892*, 2024.

Huy Phan, Oliver Y. Chên, Minh C. Tran, Philipp Koch, Alfred Mertins, and Maarten De Vos. XSleepNet: Multi-view sequential model for automatic sleep staging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5903–5915, 2022.

Sheng Qu, Xinchun Wu, Laigang Huang, Yancai Zhou, Qiangsan Sun, and Fanshuo Zeng. Diagnosis of disorders of consciousness using nonlinear feature derived EEG topographic maps via deep learning. *Scientific Reports*, 16(1):7417, 2026.

Caroline Schnakers, Audrey Vanhaudenhuyse, Joseph Giacino, Manfredi Ventura, Melanie Boly, Steve Majerus, Gustave Moonen, and Steven Laureys. Diagnostic accuracy of the vegetative and minimally conscious state: Clinical consensus versus standardized neurobehavioral assessment. *BMC Neurology*, 9(1):35, 2009.

Jacobo Diego Sitt, Jean-Rémi King, Imen El Karoui, Benjamin Rohaut, Frédéric Faugeras, Alexandre Gramfort, Laurent Cohen, Mariano Sigman, Stanislas Dehaene, and Lionel Naccache. Large scale screening of neural signatures of consciousness in patients in a vegetative or minimally conscious state. *Brain*, 137(8):2258–2270, 2014.

Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. EEG Conformer: Convolutional transformer for EEG decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2023.

Akara Supratak and Yike Guo. TinySleepNet: An efficient deep learning model for sleep stage scoring based on raw single-channel EEG. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 641–644, 2020.

Akara Supratak, Hao Dong, Chao Wu, and Yike Guo. DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(11):1998–2008, 2017.

Ling Tang, Qian Chen, Jilin Mei, Houshi Xu, Quanshi Zhang, Jing Shao, Na Zou, Xia Hu, and Dongrui Liu. What do EEG foundation models capture from human brain signals? *arXiv preprint arXiv:2605.11410*, 2026.

Jlenia Toppi, Ilaria Quattrociocchi, Angela Riccio, Mariagrazia D'Ippolito, Marta Aloisi, Emma Colamarino, Floriana Pichiorri, Febo Cincotti, Rita Formisano, and Donatella Mattia. EEG-derived markers to improve prognostic evaluation of disorders of consciousness. *IEEE Journal of Biomedical and Health Informatics*, 28(11):6674–6684, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Mário L. Vicchietti, Fernando M. Ramos, Luiz E. Betting, and Andriana S. L. O. Campanharo. Computational methods of EEG signals analysis for Alzheimer's disease classification. *Scientific Reports*, 13(1), 2023.

Fang Wang, Yu-Chu Tian, Xueying Zhang, and Fengyun Hu. An ensemble of XGBoost models for detecting disorders of consciousness in brain injuries through EEG connectivity. *Expert Systems with Applications*, 198:116778, 2022.

Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. EEGPT: Pretrained transformer for universal and reliable representation of EEG signals. In *Advances in Neural Information Processing Systems*, volume 37, 2024a.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. CBraMod: A criss-cross brain foundation model for EEG decoding. In *International Conference on Learning Representations*, 2025.

Xiaoyu Wang, Yi Yang, Geoffrey Laforge, Xueling Chen, Loretta Norton, Adrian M. Owen, Jianghong He, and Fengyu Cong. Global field time-frequency representation-based discriminative similarity analysis of passive auditory ERPs for diagnosis of disorders of consciousness. *IEEE Transactions on Biomedical Engineering*, 71(6):1820–1830, 2024b.

Yihe Wang, Nadia Mammone, Darina Petrovsky, Alexandros T. Tzallas, Francesco C. Morabito, and Xiang Zhang. ADformer: A multi-granularity transformer for EEG-based alzheimer's disease assessment. *arXiv preprint arXiv:2409.00032*, 2024c.

Yihe Wang, Nan Huang, Nadia Mammone, Marco Cecchi, and Xiang Zhang. LEAD: An EEG foundation model for alzheimer's disease detection. *arXiv preprint arXiv:2502.01678*, 2026.

Yu Wang, Yifan Gao, Xiusi Chen, Haoming Jiang, Shiyang Li, Jingfeng Yang, Qingyu Yin, Zheng Li, Xian Li, Bing Yin, et al. Memoryllm: Towards self-updatable large language models. *arXiv preprint arXiv:2402.04624*, 2024d.

Sarah Wannez, Lizette Heine, Marie Thonnard, Olivia Gosseries, Steven Laureys, et al. The repetition of behavioral assessments in diagnosis of disorders of consciousness. *Annals of Neurology*, 81(6): 883–889, 2017.

Yusuke Watanabe, Yosuke Miyazaki, Masahiro Hata, Ryohei Fukuma, Yasunori Aoki, Hiroaki Kazui, Tetsuya Araki, Daiki Taomoto, Yuta Satake, Taiki Suehiro, et al. A deep learning model for the detection of various dementia and MCI pathologies based on resting-state electroencephalography data: A retrospective multicentre study. *Neural Networks*, 171:242–250, 2024.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. TimesNet: Temporal 2d-variation modeling for general time series analysis. In *International Conference on Learning Representations*, 2023.

Fengyuan Yang, Longlin Pi, Dan Chen, Zhenhu Liang, Yong Wang, and Ye Zhang. DoC-Informer: Automated discrimination of disorders of consciousness under adaptive EEG settings. *IEEE Transactions on Biomedical Engineering*, 2026.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving mamba2 with delta rule. In *International Conference on Learning Representations*, volume 2025, pp. 29687–29707, 2025.

Ce Zhang, Young-Keun Kim, and Azim Eskandarian. EEG-Inception: An accurate and robust end-to-end neural network for EEG-based motor imagery classification. *Journal of Neural Engineering*, 18(4):046014, 2021.

Yuchen Zhou, Jiamin Wu, Zichen Ren, Zhouheng Yao, Weiheng Lu, Kunyu Peng, Qihao Zheng, Chunfeng Song, Wanli Ouyang, and Chao Gou. Csbrain: A cross-scale spatiotemporal brain foundation model for eeg decoding. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Sining Zhoubian, Dan Zhang, Evgeny Kharlamov, and Jie Tang. Memory for large language models. *arXiv preprint arXiv:2607.25380*, 2026.

APPENDIX

A SELF-COLLECTED CLINICAL EEG DATASETS

Our clinical dataset contains two linked, subject-level cohorts rather than two independent collections. MMSE-140 comprises 140 participants with a valid MMSE score, and DoC-60 is the subset of 60 participants with a clinical UWS, MCS−, MCS+, or EMCS diagnosis. The DoC-60 endpoint groups MCS−, MCS+, and EMCS as MCS/EMCS; the MMSE-140 endpoint separates the MMSE-zero and MMSE-positive groups. Figure A.1 shows the resulting class counts.

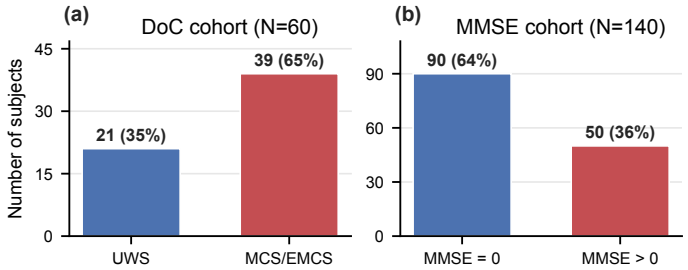


Figure A.1: Class counts and proportions for DoC-60 and MMSE-140. Panel (a) shows 21 UWS and 39 MCS/EMCS participants in DoC-60; panel (b) shows 90 participants with MMSE $= 0$ and 50 with MMSE > 0 in MMSE-140.

As shown in Figure B.1, DoC-60 includes 60 participants, comprising 40 men and 20 women. Their approximate ages span the 10–89-year age bins, with a median of 53 years.

MMSE-140 includes 140 participants, comprising 98 men and 42 women. Their approximate ages also span the 10–89-year age bins, with a median of 53 years.

B IMPLEMENTATION DETAILS

Signals are resampled to 100 Hz and divided into non-overlapping 30-second chunks. The tokenizer uses 50-sample patches with a 25-sample stride. Its two waveform convolutions have 32 output channels and kernel widths 25 and 5; the shared token dimension is 64. We use $J = 3$ EviAttention stages, each with four temporal and four electrode attention heads. The four experts use nominal windows of $r_e \in \{0.5, 1, 3, 7\}$ seconds, and the router selects two experts per token.

We first pretrain on the training set with batch size 64, learning rate 10^{-3}, and weight decay 10^{-4}. We then fine-tune on the training set with AdamW for five epochs, using batch size 128, learning rate 10^{-3}, and weight decay 10^{-4}. The temporal contrastive loss has weight 0.05 and temperature 0.2. For DoC-60, each update samples 2, 4, 8, or 16 adjacent positions from the original subject-level chunk order. For MMSE-140, each update uses 8 or 16 chunks. We train on five Tesla V100 GPUs.

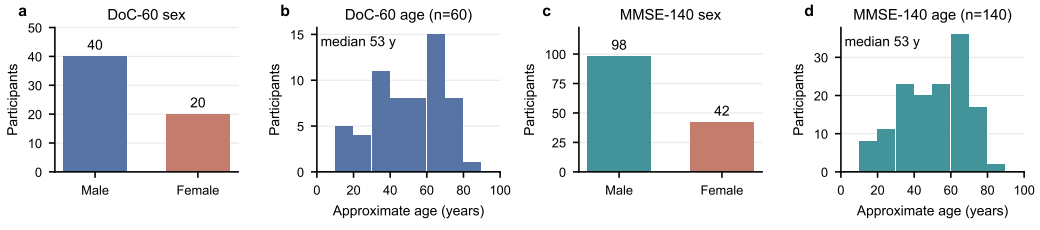


Figure B.1: Participant sex and approximate age. Panels (a–b) show DoC-60, and panels (c–d) show MMSE-140. Bars and histogram bins count participants; age medians are annotated.

C Related Work

This section further surveys recent progress on long-context processing and memory, which has attracted growing attention in natural language modeling. Researchers combine language models with implicit states and explicit memory stores to preserve and retrieve information distributed across long contexts. Zhoubian et al. organize these mechanisms by representation, update dynamics, and persistence (Zhoubian et al., 2026).

Longformer combines local windowed attention with task-dependent global attention to process long documents with linear attention complexity (Beltagy et al., 2020). Sparse Transformer factorizes the attention matrix into sparse patterns, reducing its complexity while modeling long sequences (Child et al., 2019). Mamba uses input-dependent selective state-space updates to propagate or forget information with linear scaling in sequence length (Gu & Dao, 2023). Gated DeltaNet combines a gate for rapid memory erasure with a delta rule for targeted updates (Yang et al., 2025). Eagle and Finch, the RWKV-5 and RWKV-6 models, add multi-headed matrix-valued states and dynamic recurrence to retain efficient recurrent inference (Peng et al., 2024). Switch Transformer routes each token to a selected expert, increasing model capacity through sparse conditional computation rather than a persistent memory state (Fedus et al., 2022). MEMORYLLM adds a fixed-size latent memory pool whose parameters can be updated with new text knowledge (Wang et al., 2024d). Titans combines attention over the current context with a neural long-term memory that learns to store past information at test time (Behrouz et al., 2025).

Existing work in natural language modeling demonstrates the potential of long-context dependencies and memory mechanisms. To our knowledge, we are the first to bring long-context modeling and memory to EEG-based DoC assessment.

D Mechanism Experiments

To test whether evidence-accumulation mechanisms benefit as the available EEG history grows, we construct a synthetic benchmark and compare EviFormer with representative sequence-memory models. We synthesize 16-channel EEG-like signals at 100 Hz. Each 30-second window contains slow-wave bouts lasting one to three seconds, and the two window types have high or low slow-wave occupancy. Each paired example contains the same high- and low-occupancy windows in two orders. One sequence groups windows of the same type before switching to the other type, whereas the other alternates the two types frequently. The task therefore requires learning order dependence across the observed history rather than decoding the local windows alone. We evaluate 2, 4, 6, 8, 10, 15, 20, 25, and 30 minutes of observation, corresponding to 4–60 windows, and generate the training and test waveform pools independently.

Using only training waveforms, we first train an EEG-Conformer window encoder to distinguish the two slow-wave occupancy types and then freeze it. We compare the EviFormer subject head with EEG-Conformer mean pooling, Longformer, RWKV-6, RetNet, and Samba. Each method learns a projection from the same frozen window embeddings and receives identical paired-order supervision. Models are trained on mixed observation lengths and evaluated at each duration on held-out paired examples. We report mean AUROC and sample standard deviation over three random seeds.

EEG-Conformer mean pooling remains at chance as history increases because averaging local window representations cannot model their order. In contrast, the sequence-memory models generally improve with longer histories, confirming that the benchmark requires cross-window dependence. EviFormer improves rapidly and achieves the strongest performance at 30 minutes, consistent with EviEmbedding, EviAttention, and EviMemory identifying and accumulating ordered evidence efficiently. This controlled synthetic result tests the intended mechanism; it does not establish that real DoC decisions use the same ordering process.

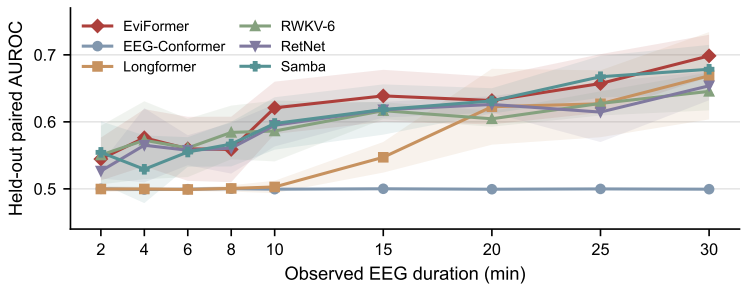


Figure D.1: Performance across observation lengths. Lines and shaded bands show mean held-out AUROC and standard deviation over three seeds, respectively.

E Statistical Analyses

Finally, we validate the relationship between EviFormer's decisions and known EEG biomarkers using expert-defined features. We test whether spindle density, slow-wave density, and the $(\delta + \theta)/(\alpha + \beta)$ slowing ratio connect the clinical phenotype, learned representation, and final decision (Tang et al., 2026). We use Mann–Whitney tests and Cliff's delta for group separation, label-adjusted partial rank correlations for alignment with the mean held-out subject probability, and nested out-of-fold ridge readouts for feature decodability from the shared pretrained embedding. Confidence intervals use participant bootstrap resampling, and Benjamini–Hochberg correction is applied within each dataset and analysis.

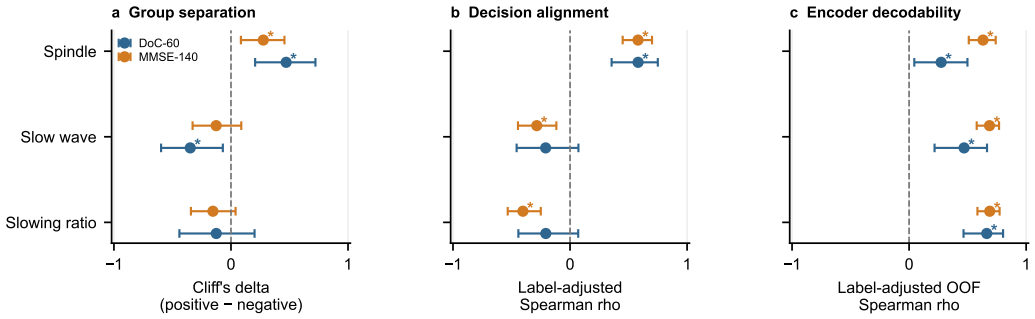


Figure E.1: Expert-defined features validate the relationship between model decisions and known EEG biomarkers.

As shown in Figure E.1, spindle density is higher in the positive class in DoC-60 (Cliff's $\delta = .471$, $q = .008$) and MMSE-140 ($\delta = .276$, $q = .021$), while MCS/EMCS has lower slow-wave density than UWS ($\delta = -.348$, $q = .042$). After controlling for diagnosis, the held-out decision remains associated with more spindles in DoC-60 ($\rho = .581$, $q < 10^{-5}$) and with more spindles, fewer slow waves, and less slowing in MMSE-140 (all $q < .01$). The shared encoder significantly represents all three descriptors in both datasets (all $q \leq .034$). Together, these analyses show that EviFormer aligns with expert knowledge that spindle and slow-wave activity are informative EEG biomarkers of consciousness while also using distributed temporal and inter-electrode structure. These associations support physiological relevance but do not establish biomarker causality.